மொழி ஆய்வில் கணிப்பொறியின் பயன்பாடு

(Use of Computer in Linguistic Research)

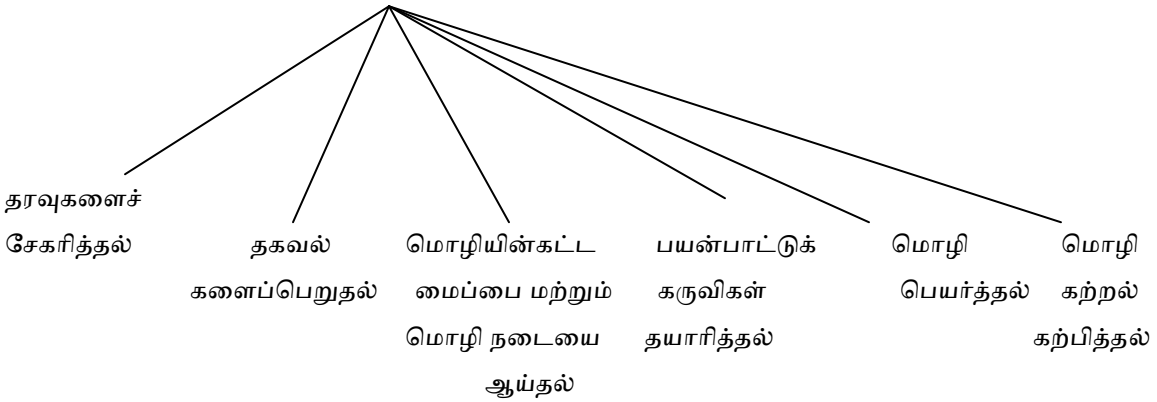
முனைவர் ச. இராசேந்திரன் *

rajushsh@gmail.com, 2005

இன்றைய அறிவியல் உலகத்தின் நாயகன் கணிப்பொறி. இருபதாம் நூற்றாண்டில் அறிமுகப்படுத்தப்பட்ட பல்வேறு பயன்பாட்டுச் சாதனங்களைப் பின் தள்ளிவிட்டு நமக்கு இன்றியமையாத பங்காளியாய் மாறியுள்ளது கணிப்பொறி. கணிக்கும் பொருளாக மட்டும் இருந்த கணிப்பொறி, தரவுகள் சேமித்து வைக்கும் கருவியாக, தகவல் தொடர்புக் கருவியாக, அத்தகவல்களைப் பயன்படுத்திக் கொள்ள உதவும் கருவியாக வளர்ந்துள்ளதோடு, நாம் தரும் தரவுகளைத் தானே ஆய்ந்து நமக்கு முடிவுகளை வெளிப்படுத்தும் கருவியாகவும் பயன்படுகின்றது. மொழிச்செயல்பாட்டிலும் கணிப்பொறியின் பங்கு நம்மை வியப்படையச் செய்யும் விதத்தில் வளர்ந்துள்ளது.

மொழித் தொடர்பான கணிப்பொறியின் பயன்பாட்டை ஆறு முக்கியக் கூறுகளாகப் பகுக்கலாம்.

மொழித்தொடர்பான கணிப்பொறியின் பயன்பாடு



முதல் இரண்டும் மொழி அல்லாத பிற நெறிகளுக்கும் பொருந்தும். மூன்றாவது பயன்பாடு பின்னர்வரும் மூன்று பயன்பாட்டிற்கும் அடிப்படையாக அமையும்.

தரவுகளை சேகரித்தல்

தற்காலக் கணிப்பொறியின் நினைவகம் மிகப்பெரிய அளவிலுள்ளது. துணை நினைவகங்களும் மிகப்பெரிய சேகரிப்பு இடத்தைக் கொண்டுள்ளன. பல கோடிக்கணக்கான சொற்களால் அமைந்த தரவுகளைக் கணிப்பொறியில் (நினைவகம், பதிவுத் தட்டு, பதிவு நாடா போன்றவற்றில்) சேகரிப்பு வைத்துத் தேவையான போது பயன்படுத்திக்கொள்ள இயலும்.

மொழிக்கான தரவுகள் பல வகைகளில் அமையலாம். சொற்களின் தொகுப்பாகவொ வாக்கியங்களின் தொகுப்பாகவோ அமையலாம். ஒரு பனுவலாக (Text) அமையலாம். (எந்த ஒரு மொழி உடலும் பனுவலாகும்.) வெவ்வேறு வகைப்பட்ட பனுவல்களில் இருந்து தெரிந்தெடுக்கப்பட்ட பகுதிகளின் தொகுப்பாக அமையலாம். இத்தகைய தரவு பெருந்தரவு (Corpus) என்று அழைக்கப்பெறும். தமிழுக்கான பெருந்தரவைச் சேகரித்து வைப்பது மிக இன்றியமையாத பணியாகும். பல இலக்கியக் கூறுகளை உள்ளடக்கிய பனுவல்கள், பல நடையியல் கூறுகளை உள்ளடக்கிய பனுவல்கள், வட்டார வழக்கு, கிளைமொழி வழக்கு, சமுதாய மொழி வழக்கு, பேச்சு வழக்கு, இலக்கிய வழக்கு போன்ற பல்வேறு மொழி வழக்குகளை உள்ளடக்கிய பனுவல்கள் எல்லாம் ஒரு சேரப் பெருந்தரவாக அமையும். இத்தகைய பெருந்தரவை இருவகையாகப் பிரிக்கலாம். ஒன்று எழுத்து வடிவில் அமையும் பெருந்தரவு (Written Corpus) மற்றொன்று பேச்சு வடிவில் அமையும் பெருந்தரவு (Speech Corpus). இத்தகைய பெருந்தரவுகள் மொழி ஆய்வுக்கும் மொழிக்கான பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பேருதவி புரியும்.

தகவல்களைப் பெறுதல்

தரவுகள் கணிப்பொறியால் ஆயப்பட்டுத் தகவல்களாகத் தரப்படும். மொழித்தொடர்பான தகவல்களைப் பல வகைப்பட்ட நினைவகத்தில் (வன் தட்டு, குறுந்தட்டு, பதிவு நாடா போன்றவற்றில்) சேமித்து வைத்துத் தேவைப்படும் போது பயன்படுத்திக்கொள்ளலாம். பல்வேறு விதமான தகவல்கள் இணையதளத்தில்

பெருமளவில் நிரம்பிக் கிடக்கின்றன. இத்தகவல்களைப் பெறுவதில் முக்கியமான அம்சம் என்னவென்றால் இயல் மொழியைப் பயன்படுத்தித் தேவையான தகவல்களைப் பெறுதலாகும். இதற்குத் தேவை இயல் மொழி இடைமுகப்பு (Natural Language Interface). இதனால் இயல் மொழி அடிப்படையிலான குறியீடுகளைக் கொண்டே தேவையான தகவல்களைப் பெற இயலும். தமிழ் மொழியின் கட்டமைப்பைக் கணிப்பொறி வழி மொழியின் எல்லா நிலைகளிலும் (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) ஆய்ந்து வைத்திருந்தால் தமிழ்மொழியைப் பயன்படுத்தித் தகவல்களைப் பெற உதவும் இயல்மொழி இடைமுகப்புகளை நிறுவ இயலும். இத்தகைய இடைமுகப்பு பல்வேறு வடிவுகளைக் கொண்டிருக்கும். தகவல் தேடும் பொறி (Search Engine) தகவல்களைப் பெற உதவும் முக்கிய கருவியாகும். தமிழுக்கு இத்தகைய தகவல் தேடும் பொறி உருவாக்குவதற்கான முயற்சிகள் மேற்கொள்ளப்பட்டும் அவை பெருமளவில் வெற்றியும் அடைந்துள்ளன.

மொழியின் கட்டமைப்பை ஆய்தல்

மொழிப்பனுவல்களைப் படிநிலை (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) அமைப்பில் வரும் மொழியலகுகளைப் பிரித்தாய்ந்து மொழியின் கட்டமைப்பைப் புரிந்துகொள்ளக் கணிப்பொறியைப் பயன்படுத்தலாம். பின்வரும் தமிழ் வாக்கியத்தை எடுத்துக்கொள்ளலாம்.

பள்ளி மாணவன் ரகுமான் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்.

பள்ளிமாணவன் ரகுமான் என்பது பெயர்த்தொடர் என்றும் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான் என்பது வினைத் தொடர் என்றும் மேலும் பள்ளி, மாணவன், மரம், பலாப்பழம், ரகுமான் என்பன பெயர்கள் என்றும், பறி என்பது வினை என்றும் பெரிய என்பது பெயரடை என்றும் கவனமாக என்பது வினையடை என்றும் இலிருந்து, ஐ என்பன வேற்றுமை உருபுகள் என்றும், இலிருந்து, த என்பது இறந்த கால இடைநிலை என்றும், ஆன் என்பது படர்க்கை

ஆண்பால் வினைமுற்று விகுதி என்றும் பகுப்பாய்வு செய்யலாம். மொழியின் கட்டமைப்பைப் பற்றிய தகவல்களைச் செயற்கை அறிவாகக் கணிப்பொறிக்கு ஊட்டினால் கணிப்பொறி மேற்கண்டது போன்ற வாக்கியங்களைப் பகுத்தாய்வு செய்து தந்து விடும்.

மொழிப் பனுவலை இடுவராலாகத் (Input) தந்து தக்க வழியமைப்புகளும் (Programs) கட்டளைகளும் தந்தால், மொழியின் கட்டமைப்பு குறித்த தகவல்களை விடுவரலாகப் (Output) பெறலாம். கணிப்பொறியைக் கொண்டு மொழிப்பனுவல்களை வாக்கியங்களாகவும் வாக்கியங்காளைத் தொடர்களாகவும் தொடர்களைச் சொற்களாகவும் சொற்களை ஒலியன்களாகவும் ஒலியன்களை ஒலிக்கூறுகளாகவும் பிரித்து ஆயலாம். மேலும் மொழி அலகுகளுக்கு இடையே உள்ள பொருண்மை உறவுகளையும் ஆயலாம். இதற்காகத் தொடரியல் ஆய்வி (Syntactic Analyzer), உருபனியல் ஆய்வி (Morphological Analyzer), ஒலியனியல் ஆய்வி (Phonological Analyzer), பொருண்மையியல் ஆய்வி (Semantic Analyzer) என்ற ஆய்விகளைத் தயாரிக்க வேண்டும். தமிழில் இத்தகைய ஆய்விகள் தயாரிக்கப்பட்டுப் பயன்பாட்டில் உள்ளன.

மொழி நடையை ஆய்தல்

மொழியின் கட்டமைப்பைக் கருத்தாடல், வாக்கியம், தொடர் சொல் என்ற அலகுகளின் அடிப்படையில் கணிப்பொறி மூலம் ஆய இயலும் என்று கண்டோம். மொழி அலகுகள் நடையியல் அடிப்படையிலான சிறப்புப் பண்புக் கூறுகளை வெளிப்படுத்தி மொழிநடை வேறுபாட்டை உணர்த்தும். கணிப்பொறியின் உதவியால் வேறுபட்ட நடையில் அமைந்த பனுவல்களை அல்லது மொழி வழக்குகளை ஆய்ந்து ஒப்பீடு செய்து அவற்றின் நடையியல் சிறப்புப் பண்புகளைத் தெரிந்து கொள்ளலாம். இத்தகைய ஆய்வு படைப்பாளியை இனங்கண்டு கொள்ளவும் படைப்பிலக்கியத்தின் காலத்தைக் கணிக்கவும் பயன்படும். கணிப்பொறி உதவியுடன் இத்தகைய ஆய்வை நுணுக்கமாகவும் விரைவாகவும் செய்து முடிக்கலாம். இத்தகைய, கணினி வழி தமிழ் மொழி நடை ஆய்வுகள் பல வெற்றிகரமாகச் செய்து முடிக்கப்பட்டுள்ளன. புள்ளியல் மொழியியல் (Statistical

Linguistics), அளவீட்டு மொழியியல் (Quantitative Linguistics) போன்ற பயன்பாட்டு மொழியியல் பாடங்களின் கோட்பாடுகளும் கொள்கைகளும் மொழியியல் நடை ஆய்வுக்கு உறுதுணை புரிகின்றன. மொழிநடையை ஆய உதவும் மென்பொருள்கள் பல கண்டு பிடிக்கப்பட்டு மொழிப் பயன்பாட்டில் உள்ளன.

பயன்பாட்டுக் கருவிகள் தயாரித்தல்

சொற்கள் தொடர்பான பயன்பாட்டுக் கருவிகள்

சொற்கள் தொடர்பான பயன்பாட்டுக் கருவிகளைத் தயாரிப்பதற்குக் கணிப்பொறியைப் பயன்படுத்தலாம். பின்வருவன இதில் அடங்கும்.

1. சொற்களை அகரவரிசைப்படுத்தல்
2. சொல்லடைவு தயாரித்தல்
3. சொல்லாட்ச்சி அகராதி மற்றும் பொருளடைவு தயாரித்தல்
4. அகராதி தயாரித்தல்
5. சொற்களஞ்சியம் தயாரித்தல்
6. எழுத்து மற்றும் இலக்கணப் பிழை திருத்தி தயாரித்தல்
7. சொல் ஆய்வி தயாரித்தல்

சொற்களை அகரவரிசைப் படுத்தல்

தமிழ்மொழியின் அகரவரிசை தொடர்பான செய்திகளையும் வழியமைப்பு ஆணைகளையும் (Program Instructions) கணிப்பொறிக்குத் தந்தால் கணிப்பொறி மிகக் குறைந்த நேரத்தில் வேண்டிய சொற்களை அகர வரிசைப் படுத்தித் தந்துவிடும். ஆங்கிலத்திற்கு இது மிகச் சாதாரண ஒன்று. தமிழைப் பொறுத்தவரையில் பலவிதமான எழுத்துருக்களும் (Fonts) அதற்குப் பின்னணியாய் அமையும் வேறுபட்ட குறியன்களும் (Codes) குறியன் எண்களும் (Code Numbers) பயன்படுத்தப்படுவதால் அகரவரிசைப் படுத்தல் எளிய செயலாக அமையவில்லை. இருப்பினும் வேறுபட்ட எழுத்துருக்களை வழியமைப்புகள் மூலம்

இணைத்து அகரவரிசைப்படுத்தும் மென்பொருள்கள் கண்டுபிடிக்கப்பட்டுப் பயன்பாட்டில் உள்ளன.

சொல்லடைவு தயாரித்தல்

சொல்லடைவு தயாரிப்பதையும் கணிப்பொறி மூலம் விரைவில் செய்துவிட இயலும். சொற்களையும் அவற்றோடு தொடர்பான செய்திகளையும் தந்தால் குறைந்த கால நேரத்தில் கணிப்பொறியின் உதவியால் சொல்லடைவு தயாரித்து விடலாம்.

எடுத்துக்காட்டாகச் சங்க இலக்கியங்களுக்குச் சொல்லடைவு தயாரிப்பதை எடுத்துக் கொள்வோம். பின்வரும் அடிகள் சங்க இலக்கியமான புறநானூற்றிலிருந்து எடுக்கப்பட்டவை.

1. இவனியாரென்குவை யாயி னிவனே (புறம் 83.1)
2. தூங்குகையா னோங்க் நடைய (புறம் 22.1)

இவ்வடிகள் சொற்களாகப் பின்வருமாறு பிரித்துத் தரப்படவேண்டும். வழியமைப்புகள் மூலம் இச்சொற்களை அகரவரிசைப்படுத்தி அவை ஒவ்வொன்றும் வரும் இடங்களைக் குறிப்பிட்டுத்தர இயலும்.

ஆயின் (புறம் 83.1)
இயார் (புறம் 83.1)
இவன் (புறம் 83.1)
என்குவை (புறம் 83.1)
ஏ (புறம் 83.1)
ஓங்கு (புறம் 22.1)
கையால் (புறம் 22.1)
தூங்கு (புறம் 22.1)
நடைய (புறம் 22.1)

ஒரு சொல் சங்க இலக்கியத்தில் எத்தனை முறை கையாளப் பெறுகின்றது அல்லது குறிப்பிட்ட ஒரு நூலில் எத்த்னை முறை கையாளப் பெறுகின்றது என்பன போன்ற தகவல்களை வழியமைப்புகள் மூலம் பெறலாம். தமிழ்ப்பல்கலைக்கழக அகராதியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்திச் சங்க இலக்கியங்களுக்கான ஒரு சொல்லடைவைத் தயாரித்துள்ளன.

சொல்லாட்சி அகராதி மற்றும் சொற்பொருளடைவு தயாரித்தல்

சொல்லாட்சி அகராதி (Concordance) ஒரு பனுவலில் இடம் பெறும் சொற்களை அகரவரிசைப்படுத்தித் தருவதோடு அச்சொற்கள் நூலில் இடம்பெறும் இடங்கள், எத்தனை முறை இடம்பெறுகின்றன என்ற கணக்கீடு, எந்தெந்தச் சூழலில் பயன்படுத்தப்படுகின்றன என்பன போன்ற செய்திகளைத் தரும். சொற்பொருளடைவு (Compendium) சொல்லாட்சி அகராதிகளினின்றும் ஒருபடி மேற்சென்று சொற்கள் வரும் இடங்களையும் சூழல்களையும் பொருள் அடிப்படையில் பிரித்துத் தரும். கணிப்பொறியைப் பொறுத்தவரையில் இரண்டையும் ஒன்றில் அடக்கித் தந்துவிடலாம்.

புறநானூற்றிலிருந்து எடுக்கப்பட்ட ஒரு சொல்லின் சொற்பொருளடைவின் மாதிரி கீழே தரப்பட்டுள்ளது.

அணி

1. அழகு
1.1 அழகு தாழ்ந்த அணி (புறம் 57.1)
1.2 அணிகொள் தேர் (புறம் 68.14)
. . .
. . .
. . .
2. அணிகலன்

2.1. இழையணிபொலிந்த . . . . . .
(புற, 81.1)

. . .

. . .

. . .

3. ஒப்பனை

3.1. ஆடு கோட்பாட்டுச் சேரலாதன் போந்தைப் பொழியின் கண் அணியல் (படி 51.9)

4. அணிமை
5. வரிசை
6. திரட்சி
7. ஆடை
8. ஒப்புமை
9. சூழ்தல்
10. பக்கரை (சேணம்)

சொல்லடைவு தயாரிக்க முன்னர்க் கூறியது போல் செய்யுளின் அடிகளைச் சொற்களாகப் பிரித்துத் தரவேண்டும். இத்துடன் சூழல் அடையாளப் படுத்தப்பட வேண்டும். வழியமைப்புகள் மூலம் நாம் சொல்லாட்சி அகராதியையோ சொற்பொருடைவையோ தயாரிக்க இயலும். தமிழ்ப்பல்கலைக்கழக் அகராதியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்திச் சங்ககால இலக்கியங்களுக்கு இத்தகைய சொல்லாட்சி அகராதியும் சொற்பொருளடைவும் தயாரித்துள்ளன.

அகராதி தாயாரித்தல்

கணினியைப் பயன்படுத்தி, பலர் உழைப்பால் பல்லாண்டுகாலம் செய்யப்படும் அகராதித் தயாரித்தலை மிகக் குறுகிய காலத்தில் குறைந்த ஆட்களின் சக்தியால் செய்து முடித்துவிடலாம். கணிப்பொறியில்

பயன்படுத்தத்தக்க விதத்தில் மின் அகராதிகளையும் தயாரிக்கலாம். இத்தகைய அகராதி கணிப்பொறி வழி மொழி கற்பித்தல், மொழிபெயர்த்தல் போன்ற செயற்பாடுகளுக்குப் பெரிதும் பயன்படும்.

அகராதி தயாரிப்பதற்கான சில மென்பொருள்கள் வெளிவந்துள்ளன. (எடுத்துக்காட்டாக Tool Box என்ற மென்பொருள் அகராதி தயாரிக்கப் பெரிதும் உதவும்.) அவற்றைப் பயன்படுத்தி அகராதிகள் தயாரிக்கலாம். DATA BASE, DATE BASE MANAGEMENT SYSTEMS போன்றவை அகராதி தயாரித்தலுக்குப் பெரிதும் உதவும். பின்வரும் எடுத்துக்காட்டு ஃபாக்ஸ்புரோ போன்ற வழிமுறைகளைப் பயன்படுத்தி அகராதி தயாரித்தலின் எளிய விளக்கமாகும்.

அகராதியின் ஒவ்வொரு சொற்பதிவிலும் தலைப்புச் சொல், எழுத்துப்பெயர்ப்பு, சொல் வகைப்பாடு, பல்பொருண்மைகள் போன்ற தகவல்கள் அடங்க வேண்டும் என்பதன் அடிப்படையில் தரவு மையத்தை உருவாக்க வேண்டும். கீழ்வரும் எடுத்துக்காட்டில் முக்காற்புள்ளிக்கு இடப் பக்கம் வருவனவற்றைப் பீல்டுகளின் பெயர்களாகவும் முக்காற்புள்ளிக்கு வலப்பக்கம் வருவனவற்றை அவற்றில் நிரப்பப்பட வேண்டிய செய்திகளாகவும் எடுத்துக்கொள்ளலாம். ஒவ்வொரு சொற்பதிவும் ஒரு பதிவாக அமையும்.

தலைப்புச் சொல்:ஏவு
எழுத்துப்பெயர்ப்பு: eevu

சொல் வகைப்பாடு: செ.பொ.குன்றா.வி.
(செயப்படுபொருள் குன்றா வினை)
பொருள் 1: தூண்டுதல்
எடுத்துக்காட்டு 1: குண்டர்களை ஏவி எங்களைப் பயமுறுத்தினான்.
பொருள் 2 : 2. கட்டவிழ்த்துவிடல்
எடுத்துக்காட்டு 2 : அடக்கு முறையை ஏவிவிட்டு மக்களின் போராட்டத்தை
அடக்கி விட இயலாது.
பொருள் 3 : அனுப்புதல், செலுத்துதல்

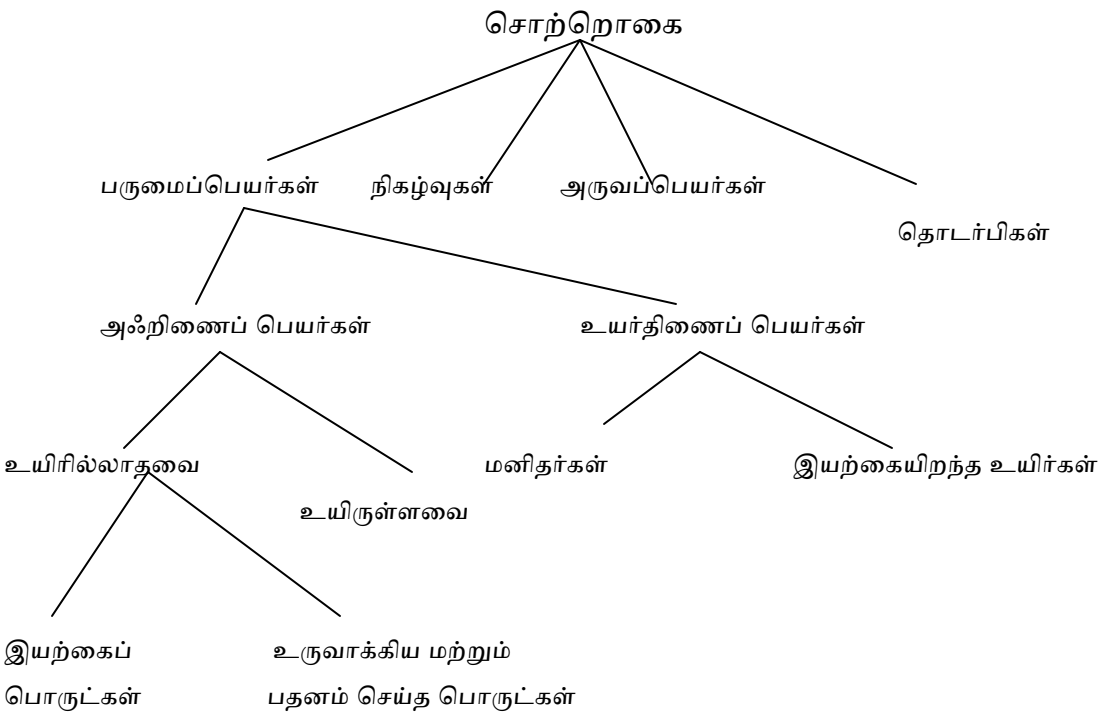
எடுத்துக்காட்டு 3 : ஏவுகணைகளை ஏவிப் பலத்த சேதம் விளைவித்தனர்.

வழியமைப்புகளைப் பயன்படுத்திச் சொற்பதிவைப் பின்வருமாறு அகராதி வடிவில் தரலாம்.

ஏவு. eevu செ. பொ. குன்ற.வி. 1. தூண்டுதல்; குண்டர்களை ஏவி எங்களைப் பயமுறுத்தினான். 2. கட்டவிழ்த்துவிடல்; அடக்கு முறையை ஏவிவிட்டு மக்களின் போராட்டத்தை அடக்கிவிட இயலாது. 3. அனுப்புதல், செலுத்துதல்; ஏவுகணைகளை ஏவிப்பலத்த சேதம் விளைவித்தனர்.

சொற்களஞ்சியம் தயாரித்தல்

சொற்களஞ்சியம் (Thesaurus) தயாரிப்பதற்குக் கணிப்பொறியைப் பயன்படுத்தலாம். ஒத்தல் (Similarity), உள்ளடங்கல் (Inclusion), மேலுறல் (Overlapping), வேறுபடல் (Disjunction) என்ற தர்க்கப் பொருண்மை உறவுகளின் அடிப்படையில் சொற்களை அவற்றின் படிநிலை அமைப்பும் பொருண்மையும் வெளிப்படும் விதத்தில் பொருண்மைக் களங்களாகப் பகுத்துத் தருவது சொற்களஞ்சியமாகும். கணிப்பொறி உதவியுடன் இச்சொற்களஞ்சியத்தைக் குறைந்த உழைப்பில் குறுகிய காலக் கட்டத்தில் தயாரித்துவிடலாம். கணிப்பொறியில் பயன்படுத்தத்தக்க விதத்தில் மின் சொற்களஞ்சியத்தை உருவாக்கலாம். இத்தகைய மின்சொற்களஞ்சியம் மொழிபெயர்த்தல், மொழி கற்பித்தல் போன்ற செயற்பாடுகளுக்குப் பெரிதும் பயனபடும். தமிழுக்கான மின் சொற்களஞ்சியம் ஒன்றைத் தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து தயாரித்துள்ளன.



இப்பாகுபாட்டின் அடிப்படையில் சொற்றொகையைப் பகுத்துப் பொருண்மை உறவுகளின் அடிப்படையில் சொற்களைத் தரவேண்டும்.

1. பருமைப் பெயர்கள்

1.1. அஃறிணைப் பெயர்கள்

1.1.1. உயிரில்லாதவை

1.1.1.1. இயற்கைப் பொருட்கள்

1.1.1.1.1. விண்வெளிப்பொருட்கள்

1.1.1.1.1.1. வானம், ஆகாயம், வான், ஆகாசம், விண், விசும்பு, அந்தரிட்சம், அந்தரம்.

1.1.1.1.1.2. கீழ்வானம்; மேல்வானம்

இது தரவு மையமாக அமையும். இதிலிருந்து வழியமைப்புகள் வழி சொற்களஞ்சியத் தகவல்களைப் பெறலாம்.

சொல்வலை தயாரித்தல்

இது ஒரு இணையதளத் தரவு மையமாகும் (Online Data Base). இது சொற்களஞ்சியம் மற்றும் அகராதி என்ற இரண்டின் பண்புகளையும் கொண்டு மிகுந்த திறன் உடையதாய் அமைந்துள்ளது. சொற்கள் தமக்கிடையே உள்ள பொருண்மை உறவுகளால் வலை போல் பின்னப்பட்டிருப்பதை வெளிப்படுத்துவதாய் இது அமையும். சொல்வலையின் மிகப் பரந்த நோக்கம் சொற்றதகவல்களைச் சொல்வடிவுகள் அடிப்படையில் அல்லாமல் சொற்பொருண்மை அடிப்படையில் ஒழுங்குபடுத்துவதாகும். இந்த அடிப்படையில் சொல்வலை அகராதியிலிருந்து வேறுபட்டுச் சொற்களஞ்சியத்தையும் உள்ளடக்கும். ரொஜஸ்ட் தெசாரஸ் போன்ற சொற்களஞ்சியங்களும் தமிழுக்கு இராசேந்திரனால் (2001) உருவாக்கப்பட்ட தமிழ்ச் சொற்களஞ்சியமும் சொல்வலை பின்னுவதற்கு முன்னோடியாய் அமையும். இவையெல்லாம் புத்தக வடிவில் அமைந்தவை. அகரவரிசையில் அமைந்த சொற்களஞ்சியங்களில் சிக்கல் என்னவென்றால் அவற்றில் தேவைக்கு அதிகமான பதிவுகள் காணப்பெறும். இரு சொற்கள் ஒருபொருள் பன்மொழிகள் என்றால் கூட அகரவரிசைக்காக இரண்டுமுறை நோக்கீடு செய்துதான் தேவையான தகவல்களைப் பெரும்பாலும் பெற இயலும். முதல்முறை அகரவரிசையில் அமைந்த சொல்லடைவு மூலம் குறிப்பிட்ட சொல் எப்பொருண்மைத் தலைப்பில் வருகிறது என்பதைத் தெரிந்துகொண்டு இரண்டாவது முறை நோக்கீடு செய்துதான் வேண்டிய தகவலைப் பெற இயலுகின்றது. இதனால் தேடும் நேரம் இரண்டு மடங்காக உயரும். ஆனால் கணிப்பொறியின் வழி தேடலை விரைவாகச் செய்ய இயலும். சொல் வலை மின்சொற்களஞ்சியத்தைவிட ஒரு படி மேலானது. தமிழ்ப் பல்கலைக்கழகமும் சென்னையிலுள்ள AUKBC Research Centre உம் இணைந்து தமிழ்ச் சொல்வலையைத் தமிழ் இணையப் பல்கலை கழகத்திற்காக உருவாக்கி உள்ளன.

எழுத்து மற்றும் இலக்கணப் பிழை திருத்தி

தமிழைத் தமிழ் எழுத்துருக்களின் உதவியால் கணிப்பொறியில் எழுத இயலும். இக்காலக் கட்டத்தில், கணிப்பொறியில் எழுதப்பட்ட தமிழ்ப்பனுவல்களில் வரும் எழுத்துப் பிழைகளையும் இலக்கணப்பிழைகளையும் கணிப்பொறியே கண்டுபிடித்துத் திருத்தம் செய்வது இன்றியமையாத தேவையாகின்றது. எழுத்து மற்றும் இலக்கணப் பிழை திருத்திட் தயாரிக்கப் பின்வரும் உட்கருவிகள் தேவையாகுகின்றன.

1. கணினித் தொடரியல் ஆய்வி
2. கணினி உருபனியல் ஆய்வி
3. கணினி ஒலியனியல்/எழுத்து ஆய்வி
4. மின் அகராதி

தமிழ்ப்பல்கலைக் கழக மொழியியல் துறையில் இவ்வாய்வு மேற்கொள்ளப்பட்டுள்ளது. தமிழுக்கு எழுத்துப்பிழை திருத்தி ஓரளவுக்குச் செம்மையாக வேலை செய்யும் வகையில் உருவாக்கப்பட்டுள்ளது. இலக்கணப் பிழைதிருத்தி முழுவதுமாக உருவாக்கப்பெறவில்லை. அதற்கான ஆய்வுகள் மேற்கொள்ளப்பட்டு முயற்சிகள் எடுக்கப்பட்டுள்ளன.

தமிழ்ச் சொல்லாய்வி தயாரித்தல்

தமிழுக்கான சொல்லாய்விகள் பல இன்று வெளி வந்துள்ளன. அவை தமிழ்ப்பனுவல்களை உள்ளீடு செய்து வடிவமைத்து வெளியீடாகப் பெறும் நுட்பங்களைக் கொண்டுள்ளன. Word, Word Star போன்ற மென்பொருள்களில் கிடைக்கும் எல்லா வசதிகளும் இவற்றிலும் உள்ளன. சொற்களஞ்சியம், எழுத்து மற்றும் இலக்கணப் பிழைதிருத்தி என்பன மேலும் மேம்படுத்தப்படவேண்டும். பல தமிழ்ச் சொல் சொல்லாய்விகளில் இவ்வசதிகள் இல்லை.

எழுத்துப் பெயர்ப்பு செய்தல்

ஒரு மொழியின் எழுத்துக்களால் எழுதப்பட்ட பிற மொழிகளின் எழுத்துக்களைப் பயன்படுத்திப் படிக்கத்தக்க முறையில் பெயர்ப்பு செய்யும் வசதிமுறைகள் கணிப்பொறியில் வந்துவிட்டன. இதனால் தமிழ் எழுத்தில் எழுதப்பட்ட பனுவல்களை ரோமன் எழுத்தில் படிக்கவோ ரோமன் எழுத்தில் எழுதப்பட்ட பனுவல்களைத் தமிழ் எழுத்தில் படிக்கவோ இயலும்.

ஒலிபெயர்ப்பு செய்தல்

கணிப்பொறியில் எழுதப்பட்ட பனுவலை ஒலிப்பு வடிவில் வரும்படி செய்தலும் முக்கியமான ஒரு ஆய்வாகும். இத்தகைய ஒலிபெயர்ப்பிகளைப் பனுவலிலிருந்து பேச்சாகப் புரிந்துகொள்ள உதவும் கருவியை Text to Speech Recognizer என்பர். தமிழில் இதற்கான முயற்சிகள் தொடங்கப்பட்டு ஓரளவுக்கு வெற்றியும் கிட்டியுள்ளது.

பேச்சுக்கூட்டிணைப்பான் மற்றும் பேச்சு அறிவான் தயாரித்தல்

பேச்சுக்கூட்டிணைப்பான் (Speech Synthesizer) மற்றும் பேச்சு அறிவான் (Speech Recognizer) தயார்செய்வது கணிப்பொறியை நேரடிமொழிப் பரிமாற்றக் கருவியாக மாற்றவும் ரோபட் போன்றவற்றை இயல்மொழியின் மூலமும் இயக்கவும் முடியும். இத்தகைய கருவிகள் தயாரிப்பதற்கான முயற்சிகள் மேற்கொள்ளப்பட்டு நடந்து வருகின்றன.

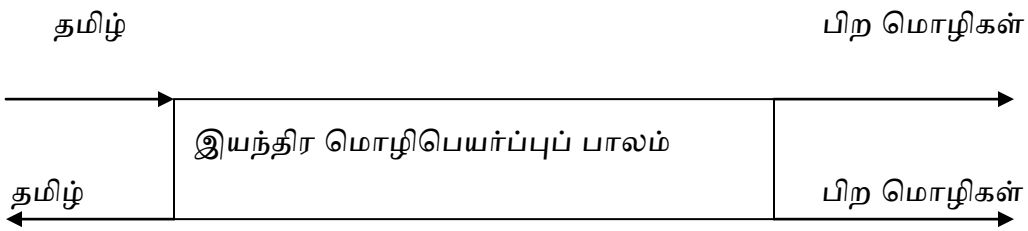
பார்வை எழுத்தறிவான் தயாரித்தல்

அச்சிடப்பட்ட பனுவல்களை அப்படியே கணினிக்கு எழுத்துக்களாகத் தானாகவே உள்ளீடு செய்ய இயலும் கருவியைப் பார்வை எழுத்தறிவான் (Optical Character Recognitior) என்பர். இக்கருவி தட்டச்சு முறைப்படி சாவிப்பலகை வழி பனுவல்களை உள்ளீடு செய்ய ஆகும் கால விரையத்தை இல்லாமல் செய்கின்றது.

ஆங்கிலத்திற்கு இத்தகைய கருவி உருவாக்கப்பட்டுப் பயன்படுத்தப்படுகின்றது. ஆனால் தமிழுக்கு அதன் எழுத்துக்களின் கலவைத் தன்மை காரணமாக இத்தகைய கருவியை உருவாக்குவது கடினமான செயலாகும். இருப்பினும் தமிழுக்கு இத்தகைய பார்வை எழுத்தறிவான் உருவாக்கும் முயற்சிகள் எடுக்கப்பட்டு ஓரளவுக்கு வெற்றியும் கிட்டியுள்ளது.

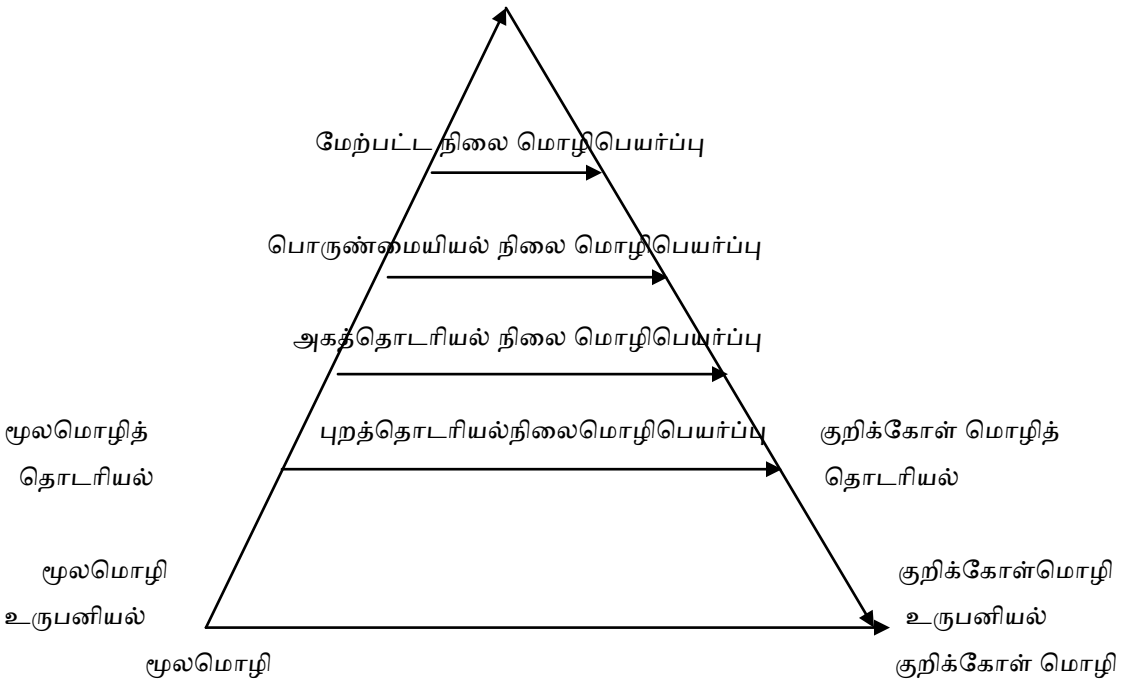
மொழிபெயர்ப்புச் செய்தல்

ஒரு மொழியிலுள்ள செய்திகளைப் பிற மொழிக்கு மாற்றித்தருவதை மொழிபெயர்ப்பு என்று சுருக்கமாய்க் கூறலாம். மொழி பெயர்ப்பு செய்வதற்கு மனித மூளைதான் தலைசிறந்த கருவி என்றாலும், மனித மொழிபெயர்ப்பில் ஏற்படும் கால தாமதத்தைத் தவிர்க்க இயந்திர மொழிபெயர்ப்பைக் கைக்கொள்ளலாம். இயந்திர மொழிபெயர்ப்பு செய்வதற்கான இயந்திரம் கணிப்பொறிதான். தமிழிலிருந்து பிற மொழிகளுக்கு மொழிபெய்ர்ப்பதற்கும் பிற மொழிகளிலிருந்து தமிழ்மொழிக்கு மொழிபெயர்ப்பதற்கும் ஒரு இயந்திர மொழிபெயர்ப்புப் பாலம் அமைத்தால் அது மொழிப் பரிமாற்றத் தடங்கலை உடைத்தெறியும்.



மூல மொழி மற்றும் குறிக்கோள் மொழி இவற்றின் கட்டமைப்புக்குத் தகுந்தவாறு இயந்திர மொழிபெயர்ப்பின் தன்மை மாறும். அதிக ஒற்றுமையுள்ள இரு மொழிகளைச் சொல் அல்லது உருபனியல் நிலையிலேயே நேரடியாக மொழி பெயர்க்கலாம். மொழிகளின் கட்டமைப்பு வேறுபாட்டிற்குத் தகுந்தவாறு, புறத்தொடர் நிலையிலோ அகத்தொடரியல் நிலையிலோ பொருண்மையியல் நிலையிலோ மொழிபெயர்க்கலாம்.

எந்த அளவிற்கு இயந்திர மொழிபெயர்ப்பில் மனித உழைப்புப் பயன்படுகிறது என்பதன் அடிப்படையில் இயந்திர மொழிபெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்.



எந்த அளவிற்கு இயந்திர மொழிபெயர்ப்பில் மனித உழைப்பு பயன்படுகிறது என்பதன் அடிப்படையில் இயந்திர மொழி பெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்.

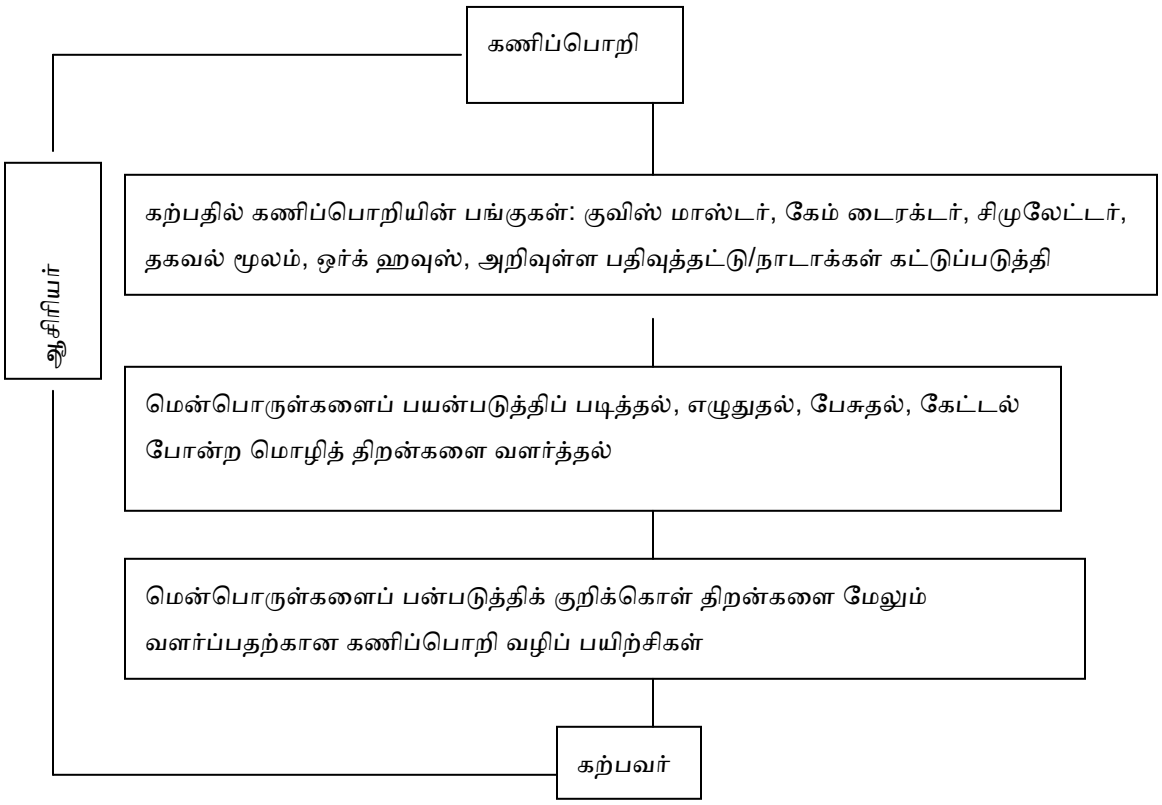
1. மனித உதவியுடன் நடைபெறும் இயந்திர மொழிபெயர்ப்பு
2. கணிப்பொறி உதவியுடன் நடைபெறும் இயந்திர மொழி பெயர்ப்பு
3. கணிப்பொறி தானே மொழிபெயர்க்கும் இயந்திர மொழி பெயர்ப்பு

இயந்திர மொழிபெயர்ப்புக்குப் பொருண்மையியல் ஆய்வி மற்றும் ஆக்கி, தொடரியல் ஆய்வி மற்றும் ஆக்கி, மொழிபெயர்ப்பு அகராதி என்பன போன்ற துணைக்கருவிகளைத் தயாரித்தல் வேண்டும்.

மொழிகற்றல்/கற்பித்தல்

மொழியைக் கற்பதற்கும் கற்பிப்பதற்கும் கணிப்பொறியைப் பயன்படுத்தலாம். வெறும் எண், எழுத்து இவற்றை மட்டும் கையாண்டு வந்த கணிப்பொறி எண், எழுத்து, படம் (Graphics), ஒலி (Audio), ஒளி (Video) அசையும் உருவங்கள் (Animation) எனப்பலவற்றை ஏற்றுக்கொண்டுள்ளது. இப்படி வேறுபட்ட பன்முகத் தொடர்புகள் கொண்ட ஒருங்கிணைப்பு மல்டிமீடியா (Multimedia) எனப்படும். மல்டிமீடியாவைப் பயன்படுத்தித் தமிழ்க்கற்றலையும் கற்பித்தலையும் ஆர்வத்தைத் தூண்டும் செயலாக மாற்றலாம்.

கணிப்பொறி அடிப்படையிலான கணிப்புச் செயல்களைக் கல்வி மற்றும் உளவியல் அடிப்படையிலான கற்றலுக்கும் கற்பித்தலுக்கும் பயன்படும் படி செய்வதுதான் கணிப்பொறி மூலம் கற்றல்/கற்பித்தல் மென்பொருள்களைத் தயாரித்தலின் நோக்கமாகும். இங்குக் கணிப்பொறி, கற்றல் அடிப்படையிலான மொழிச்செயல்பாடுகளைச் செய்வதற்குப் பல பங்குகள் வகிக்கும். எடுத்துக்காட்டாக குவிஸ்மாஸ்டர், கேம் டைரக்டர், சிமுலேட்டர், ஒர்க் ஹவுஸ், தகவல் மூலம் கட்டுப்படுத்தி போன்ற பல பங்குகளைக் கணிப்பொறி வகிக்கும். பின்வரும் வரைபடம் கணிப்பொறியின் உதவியுடன் கூடிய கற்றல்/கற்பித்தல் செயற்பாட்டைத் தெளிவுபடுத்தும்.


கணிப்பொறி
ஆசிரியர்
கற்பதில் கணிப்பொறியின் பங்குகள்: குவிஸ் மாஸ்டர், கேம் டைரக்டர், சிமுலேட்டர், தகவல் மூலம், ஒர்க் ஹவுஸ், அறிவுள்ள பதிவுத்தட்டு/நாடாக்கள் கட்டுப்படுத்தி
மென்பொருள்களைப் பயன்படுத்திப் படித்தல், எழுதுதல், பேசுதல், கேட்டல் போன்ற மொழித் திறன்களை வளர்த்தல்
மென்பொருள்களைப் பன்படுத்திக் குறிக்கொள் திறன்களை மேலும் வளர்ப்பதற்கான கணிப்பொறி வழிப் பயிற்சிகள்
கற்பவர்

மேற்கண்டவாறு கணிப்பொறியைப் பயன்படுத்தி மொழித்தொடர்பான செயற்பாடுகளைச் செய்ய மொழியியலாரும் கணிப்பொறியியலாரும் இணைந்து செயலாற்ற வேண்டும். ஒருவருடைய செயற்பாட்டை மற்றவர் புரிந்துகொண்டு செயலாற்றினால் மொழித்தொடர்பான கணிப்பொறிக் கருவிகளை, மென்பொருள்களை எளிதில் உருவாக்கிப் பயன்படுத்தலாம்.

* இணைப்பேராசிரியர் மற்றும் துறைத்தலைவர், மொழியியல் துறை, தமிழ்ப்பல்கலைக்கழகம், தஞ்சாவூர்-613 005.

துணை நூல்கள்

இராஜேந்திரன், ச., 2001 *தற்காலத் தமிழ்ச்சொற்களஞ்சியம்*, தமிழ்ப்பல்கலைக்கழகம், தஞ்சாவூர்.

" 1991 'தற்காலத் தமிழ்ச் சொற்களஞ்சியம் (பாகம் 1),' தமிழ்க்கலை, 9.1. -4:55-76.

" 1999 *'பொருட்புல வகைப்பாடும் சொற்களஞ்சியமும்,'* புலமை, 25.2:47-66.

" 2002 '*இணையம் வழி அயல்நாடுகளில் தமிழ் கற்றலை வளர்த்தல்*, 'தமிழ் ஆய்வில் இன்றைய போக்குகள், 75-81, திருமதி செண்பகம் சுப்பையா அறக்கட்டளை, தஞ்சாவூர்.

இராஜேந்திரன், ச., 2004 *தமிழ்மின்சொற்களஞ்சியம்*
பாஸ்கரன், ச. 2004 *(ஆய்வேடு),* தமிழ்ப்பல்கலைக்கழகம், தஞ்சாவூர்.

Kamakshi, S., & Rajendran, S.,	2004	"'Preliminaries to the Preparation of a Machine Aid to Translate Linguistic Texts in English into Tamil', Dravidian Linguistic Association, Thiruvananthapuram.
Rajendran, S.,	1995	' A Machine Bridge for Tamil' (To be published in Proceedings of English World Tamil Conference)
"	1997	'Grammatical Formalism and Computational Analysis of Nominal Compounds in Tamil', South Asian Language Review, Vol 7, No.1., 27-46.
"	2003	'Prerequisite for the Preparation of an Electronic Thesaurus for a Text Processor in Indian Languages', Languages in India 3:1, www.languageinindia.com.
"	2003	'Dravidian WordNet: A proposal,' In R.M. Sundaram et al (edited), 467-497, Facets of Language, Thanjavur.
"	2003	'Creating Generative Lexicon from MRDs: Tamil Experience', In: Rajeev Sangal et al (edited), 83-91, Recent Advances in Natural Language Processing, Proceedings of the International Conference of Natural Language Processing (ICON 2003),
Rajendran, S., et al	2000	'Tamil WordNet' Proceedings of the First Global WordNet Conference, CIIL, Mysore.
"	2003	'Computational Morphology of Verbal Complex', In:B Ramakrishna Reddy (edited), Word Structure in Dravidian, University, Kuppam & Language in India, 3:4, www.languageinindia.com. April, 2003.

♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠